
# Provenance for Earth Science (PROV-ES) Working Group

ESDSWG 2015 Meeting
GSFC, Baltimore, MD
Tuesday, March 24, 2015

Technical Chairs: Hook Hua
ESDIS POC: Hampapuram Ramapriyan

# PROV-ES 2014 Action Plan

## Mission Statement

*We will create an interoperable provenance specification for use in Earth Science Data Systems to enable:*

- *Capturing the increasing amount of **contextual** processing information of Earth Science Data Records (ESDRs).*
- *Improving **transparency, understanding, and trust** of ESDRs .*
- *Providing an **interoperable** representation of provenance for NASA EOS data products.*
- *Reducing the barrier of entry to **infusing provenance** capture and consumption in Earth Science Data Systems.*

## Stakeholders

- *Data Producers* - NASA Missions and projects
- *Data Processing Systems* – Enable capturing and storing provenance
- *Data Archive Centers* – Enable sharing and seamlessly linking of provenance between data centers.
- *Data Users* – Consuming provenance of data products.
- *Tool Makers* – Enable provenance exploration, visualization, and analysis.

## Approach

- Refine relevant provenance **use cases** for Earth Science Data Systems.
- Refine PROV-ES **content model** based on updated use cases.
- Create an **encoding** (JSON) of the PROV-ES content model
- Prototyping PROV-ES **implementations**
  - 2-way synergy between an ESDSWG working group and infused projects
  - Capture lessons learned.
- Develop a **"How-to" guide** for PROV-ES
  - Recommendations on how to produce, consume, and interpret provenance for ESDRs.
- Coordinate with **ESIP Federation**. E.g. Semantic Web cluster and Data Stewardship Committee

## Deliverables

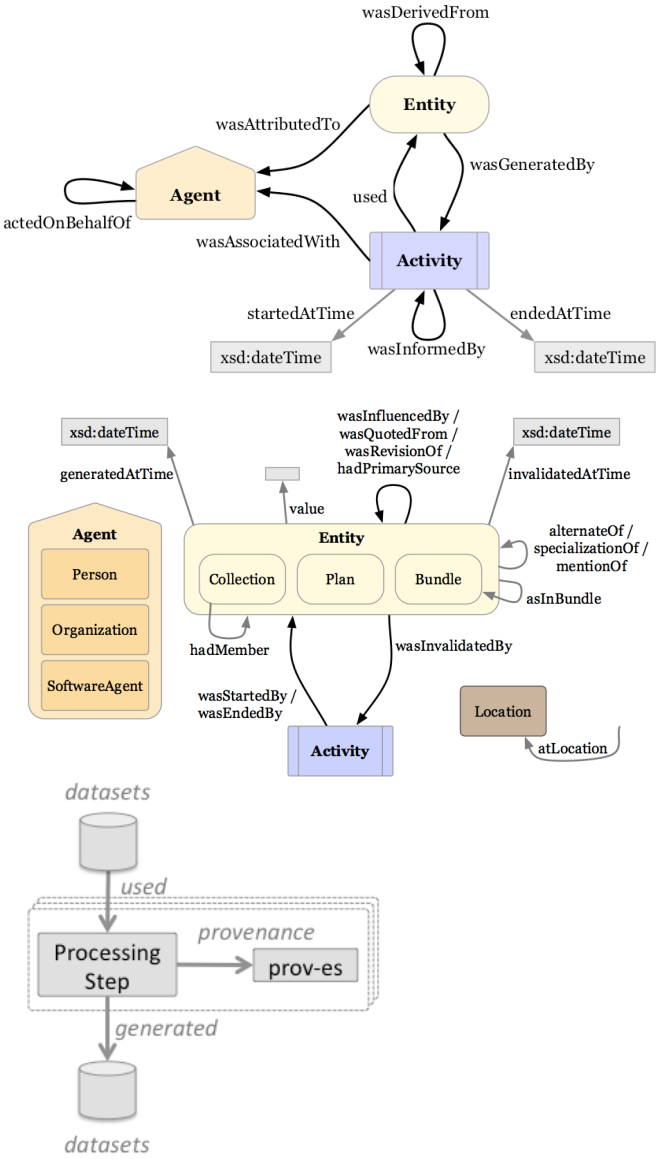
- Refined provenance **content model** for earth science data processing of ESDRs.
- Create a provenance **encoding** of the PROV-ES content model.
- Initial **implementations** of PROV-ES and **lessons learned**.
- **How-to guide** for PROV-ES: recommendations on how to produce, consume, and interpret provenance for ESDRs.

# Summary of PROV-ES


- Started as an ESDSWG 2013 Working Group
  - Developed first OPM/PROV approach from ACCESS 2009 project
  - 2013 WG developed initial model and feasibility in PROV-N encodings
- 2014 WG focused on taking 2013 results and developing working implementations
- An **interoperable** provenance specification for use in Earth Science Data Systems to enable:
  - Capturing the increasing amount of contextual processing information of Earth Science Data Records (ESDRs).
  - Improving transparency, understanding, and trust of ESDRs .
  - Providing an interoperable representation of provenance for NASA EOS data products.
  - **Reducing the barrier of entry** to **infusing** provenance capture and consumption in Earth Science Data Systems.
- Employed additional use cases from ESIP, NASA-funded projects, and USGCRP/GCIS
- Initial specification and implementations employed
- Automatic PROV-ES generation infused into initial NASA data systems
- Faceted search interface to explore PROV-ES records

# PROV-ES Implementation

- **Interoperability** with W3C PROV specification
  - Maximize reuse of available tools interoperable with PROV
    - Examples include (but not limited to): ProvToolbox (Java), Prov (Python), Prov-O-Viz (Javascript)
- Support **additional attributes** to W3C PROV to encode PROV-ES concepts
  - PROV Earth Science mapping to existing W3C PROV terms
- Support domain-specific extended concepts
- Provenance capture **instrumented** into processing steps
  - Entities, activities, and agents are kept loosely-coupled to enable more versatile provenance curation
  - E.g. instrumentation Python API for PROV-ES
- Ease **infusion** through (re)use of existing interoperable W3C PROV tools



# PROV-ES Concepts and Properties

| ES concept | Base PROV concept | Additional Attributes | | | Notes |
|---|---|---|---|---|---|
| | | key name | value type | cardinality | |
| eos:dataset | prov:Entity | | | | |
| | | *prov:type* | *eos:dataset* | | |
| | | dcterms:title | xsd:string | | human-readable name |
| | | info:doi | doi | [0..1] | leave it blank for now until we get a DOI |
| | | gcis:accessURL | xsd:anyURI | [0..*] | Could be any kind of URL that gives access to a distribution of the dataset. E.g. landing page, download, feed URL, SPARQL endpoint. Use when your catalog does not have information on which it is or when it is definitely not a download. |
| | | gcis:downloadURL | xsd:anyURI | [0..*] | This is a direct link to a downloadable file in a given format. E.g. CSV file or RDF file. The format is described by the distribution's dc:format and/or dcat:mediaType. |
| | | gcis:sourceInstrument | eos:instrument | [0..*] | A dataset was collected by using one or more instruments. |
| | | eos:level | xsd:string | [0..1] | http://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/ |
| | | eos:version | xsd:string | [0..1] | |
| eos:platform | prov:Entity | | | | |
| | | *prov:type* | *eos:platform* | | |
| | | dcterms:title | xsd:string | | human-readable name |
| | | gcis:hasInstrument | eos:instrument | [0..*] | |
| eos:instrument | prov:Entity | | | | |
| | | *prov:type* | *eos:instrument* | | |
| | | dcterms:title | xsd:string | | human-readable name |
| | | gcis:inPlatform | eos:platform | | |
| | | gcis:hasSensor | eos:sensor | [0..*] | |
| | | gcis:hasGoverningOrganization | prov:Organization | [0..*] | |
| eos:sensor | prov:Entity | | | | |
| | | *prov:type* | *eos:sensor* | | |
| | | dcterms:title | xsd:string | | human-readable name |
| | | gcis:inInstrument | eos:instrument | | |
| eos:software | prov:Entity | | | | |
| | | *prov:type* | *eos:software* | | |
| | | dcterms:title | xsd:string | | human-readable name |
| | | gcis:implements | eos:algorithm | [0..*] | |
| | | eos:version | xsd:string | [0..1] | |
| eos:algorithm | prov:Entity | | | | |
| | | *prov:type* | *eos:algorithm* | | |
| | | dcterms:title | xsd:string | | human-readable name |
| | | gcis:implementedIn | eos:software | [0..*] | |
| | | eos:describedBy | bibo:Document | [0..*] | |
| | | eos:version | xsd:string | [0..1] | |
| eos:document | prov:Entity | | | | |
| | | *prov:type* | *bibo:Document* | | |
| | | dcterms:title | xsd:string | | human-readable name |
| | | info:doi | doi | [0..1] | |
| | | gcis:downloadURL | xsd:anyURI | [0..*] | |
| | | eos:version | xsd:string | [0..1] | |
| eos:processStep | prov:Activity | | | | |
| | | *prov:type* | *eos:processStep* | | 3 |
| | | dcterms:title | xsd:string | | human-readable name |
| | | eos:usesSoftware | eos:software | [0..*] | |
| | | prov:wasAssociatedWith | prov:SoftwareAgent | [0..1] | A software agent is running software. |
| | | eos:runTimeContext | eos:runTimeContext | [0..1] | |
| | | prov:used | eos:dataset | [0..*] | |
| | | prov:generated | eos:dataset | [0..*] | |
| eos:softwareAgent | prov:SoftwareAgent | | | | |
| | | *prov:type* | *prov:SoftwareAgent* | | |
| | | dcterms:title | xsd:string | | human-readable name |
| | | id | xsd:string | [0..1] | |
| | | host | xsd:string | [0..1] | |
| eos:publication | prov:Entity | | | | A bounded physical representation of a body of information designed with the capacity (and usually intent) to communicate, and which has been made available to the public. |
| | | *prov:type* | *gcis:publication* | | |

# PROV-ES Infusion

- AIST 2011 on **geodetic InSAR processing**
  - PROV-ES production in NASA AIST's Advanced Rapid Imaging & Analysis for Monitoring Hazards (ARIA-MH) Hybrid Cloud Science Data System (HySDS)
- MEaSUREs 2012 on **A-Train atmospheric science data fusion**
  - PROV-ES production in NASA MEaSUREs's "A Multi-Sensor Water Vapor, Temperature and Cloud Climate Data Record"
- Additional Use Case
  - **Mapping NCA3 provenance from GCIS** into PROV-ES

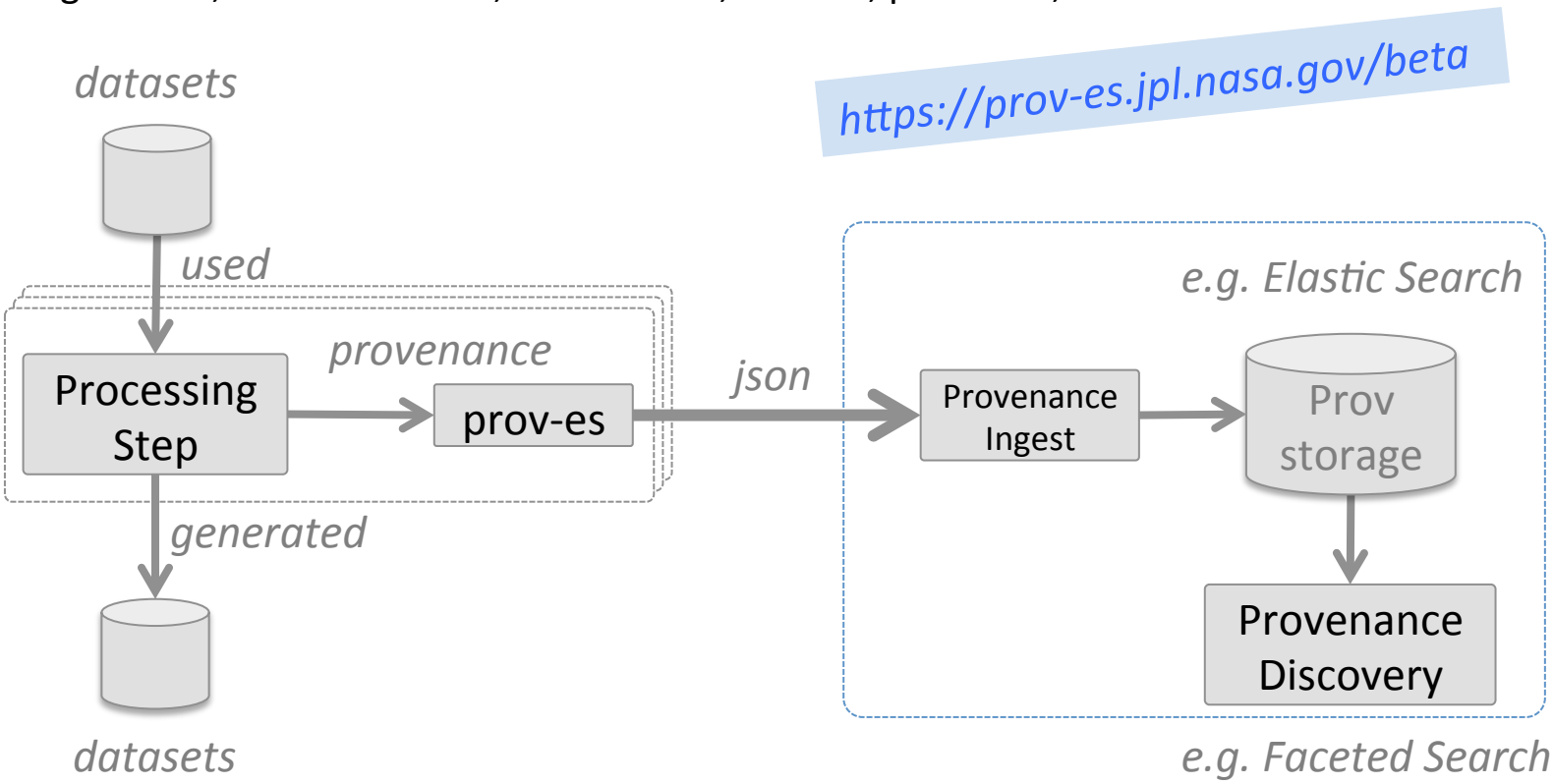
# PROV-ES JSON Examples

```
{
  "used": {
    "hysds:25a3a23d-
      "prov:time": "
      "prov:entity":
      "prov:activity"
    },
  },
 "entity": {
    "eos:ISCE": {
      "dcterms:title"
      "eos:version":
      "prov:type": "e
      "gcis:implement
    },
    "eos:SRTM": {
      "gcis:hasSensor
      "gcis:inPlatfor
      "prov:type": "e
      "gcis:hasGoverr
    },
  },
  "activity": {
    "hysds:380674f9-2
      "eos:runTimeCon
      "prov:wasAssoci
      "dcterms:title"
      "prov:startTime
      "prov:type": "e
      "eos:usesSoftwa
      "hysds:job_id":
CSKS4_RAW_HI_21_HH_RI
      "prov:endTime":
    }
  },
  "wasGeneratedBy":
    "hysds:52a02b04-5
      "prov:time": "2
      "prov:entity":
      "prov:activity"
    }
  }
}
```

```
{
"entity": {
    "gcis:chapter-our-changing-climate": {
      "dcterms:title": "Our Changing Climate",
      "gcis:accessURL": "http://nca2014.globalchange.gov/report/our-changing-climate/introduction",
      "prov:type": "gcis:Chapter"
    },
    "gcis:report-nca3": {
      "dcterms:title": "Climate Change Impacts in the United States: The Third National Climate Assessment",
      "gcis:accessURL": "http://nca2014.globalchange.gov",
      "prov:type": "gcis:Report"
    },
    "gcis:image-006264b2-8594-4e09-9430-a1d0cc056ff7": {
      "dcterms:title": "Projected Precipitation Change by Season (Winter)",
      "gcis:accessURL": "http://data.globalchange.gov/image/006264b2-8594-4e09-9430-a1d0cc056ff7",
      "prov:type": "gcis:Image"
    },
    "gcis:figure-projected-precipitation-change-by-season": {
      "dcterms:title": "Projected Precipitation Change by Season",
      "gcis:accessURL": "/report/nca3/chapter/our-changing-climate/figure/projected-precipitation-change-by-season",
      "prov:type": "gcis:Figure"
    },
    "gcis:dataset-nca3-cmip3-r201205": {
      "dcterms:title": "dataset nca3-cmip3-r201205",
      "gcis:accessURL": "http://data.globalchange.gov/dataset/nca3-cmip3-r201205",
      "prov:type": "eos:dataset"
    }
  },
  "activity": {
    "gcis:activity-006264b2-nca3-cmip3-r201205-process": {
      "prov:startTime": "2013-07-24T00:00:00",
      "prov:endTime": "2013-07-24T00:00:00",
      "prov:wasAssociatedWith": "gcis:contributor-2522"
    }
  },
  "wasGeneratedBy": {
    "_:8a24a829-93b6-5390-be83-942a89608d1f": {
      "prov:entity": "gcis:image-006264b2-8594-4e09-9430-a1d0cc056ff7",
      "prov:time": "2013-07-24T00:00:00",
      "prov:activity": "gcis:activity-006264b2-nca3-cmip3-r201205-process"
    }
  }
}
```

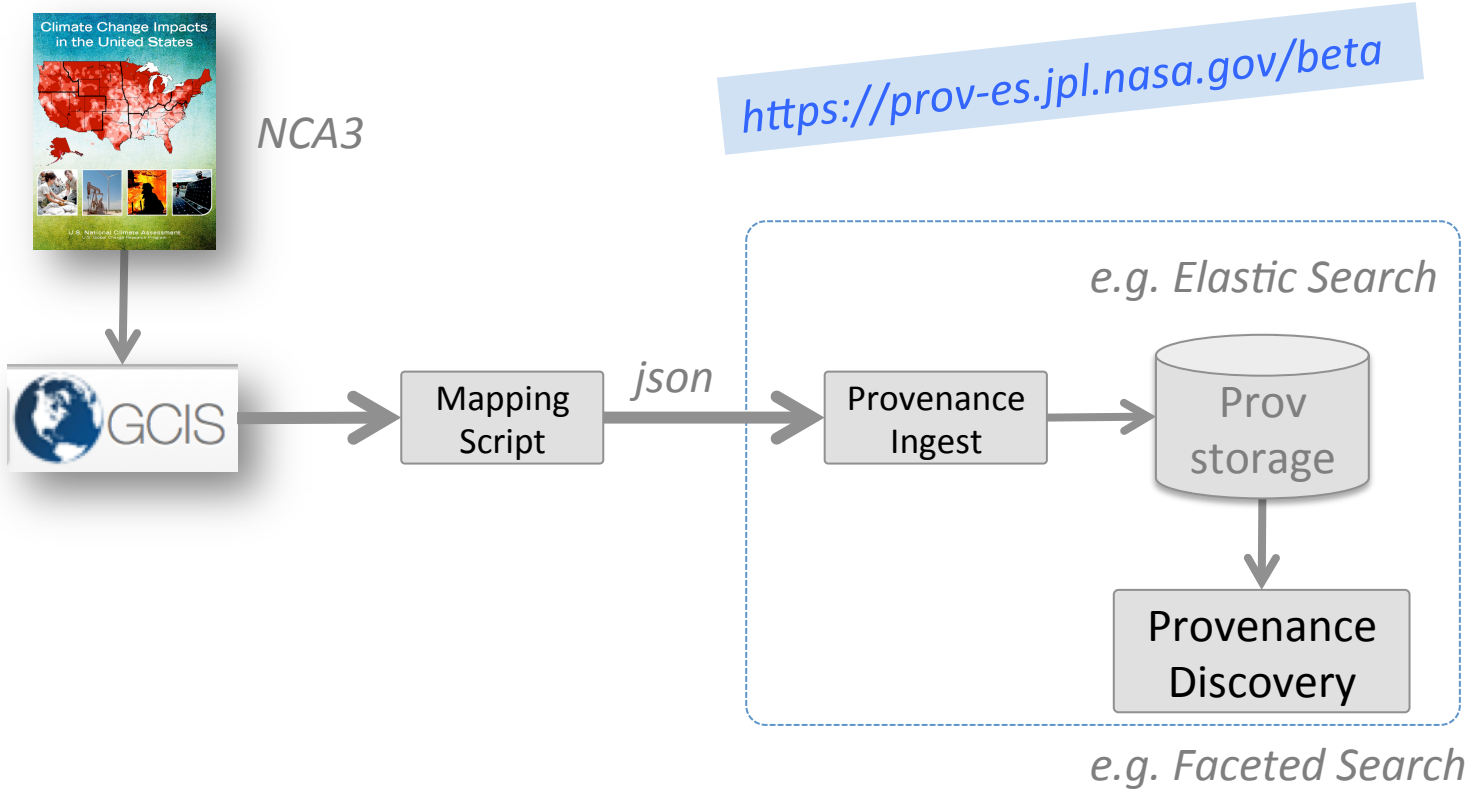
# From Instrumentation to Faceted Search

- PROV-ES “log lines” generated in JSON
- Distributed PROV-ES traces crawled and ingested into Elastic Search datastore
- Support discovery of multi-dimensionality of provenance through **Faceted Search**
  - Enables drilling down across end-to-end provenance concepts such as datasets, processing steps, algorithms, documentation, instruments, sensors, platforms, etc.

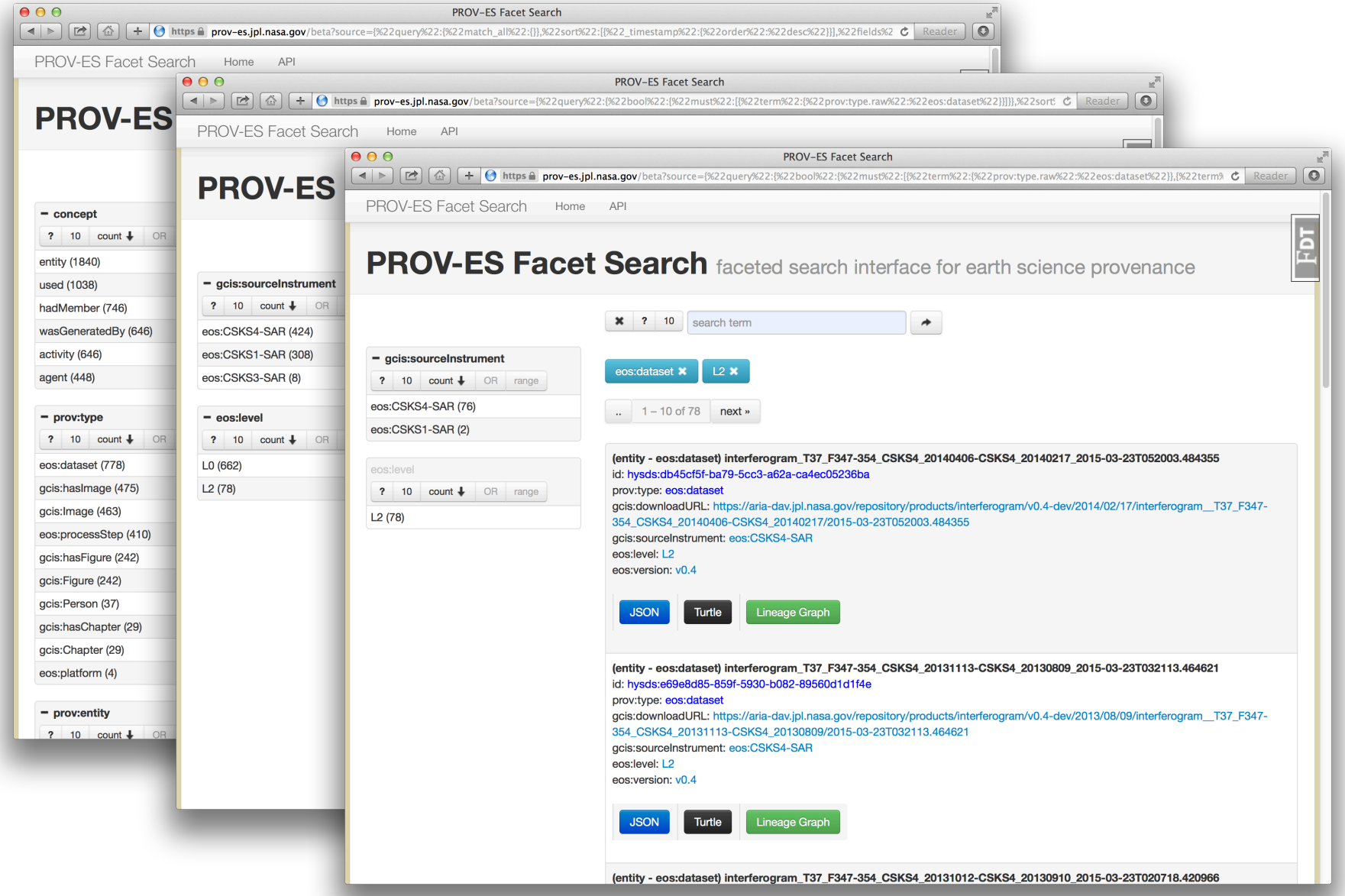


# Mapping From Other Sources

- Support Faceted Search of NCA3 Report provenance by mapping from GCIS into PROV-ES



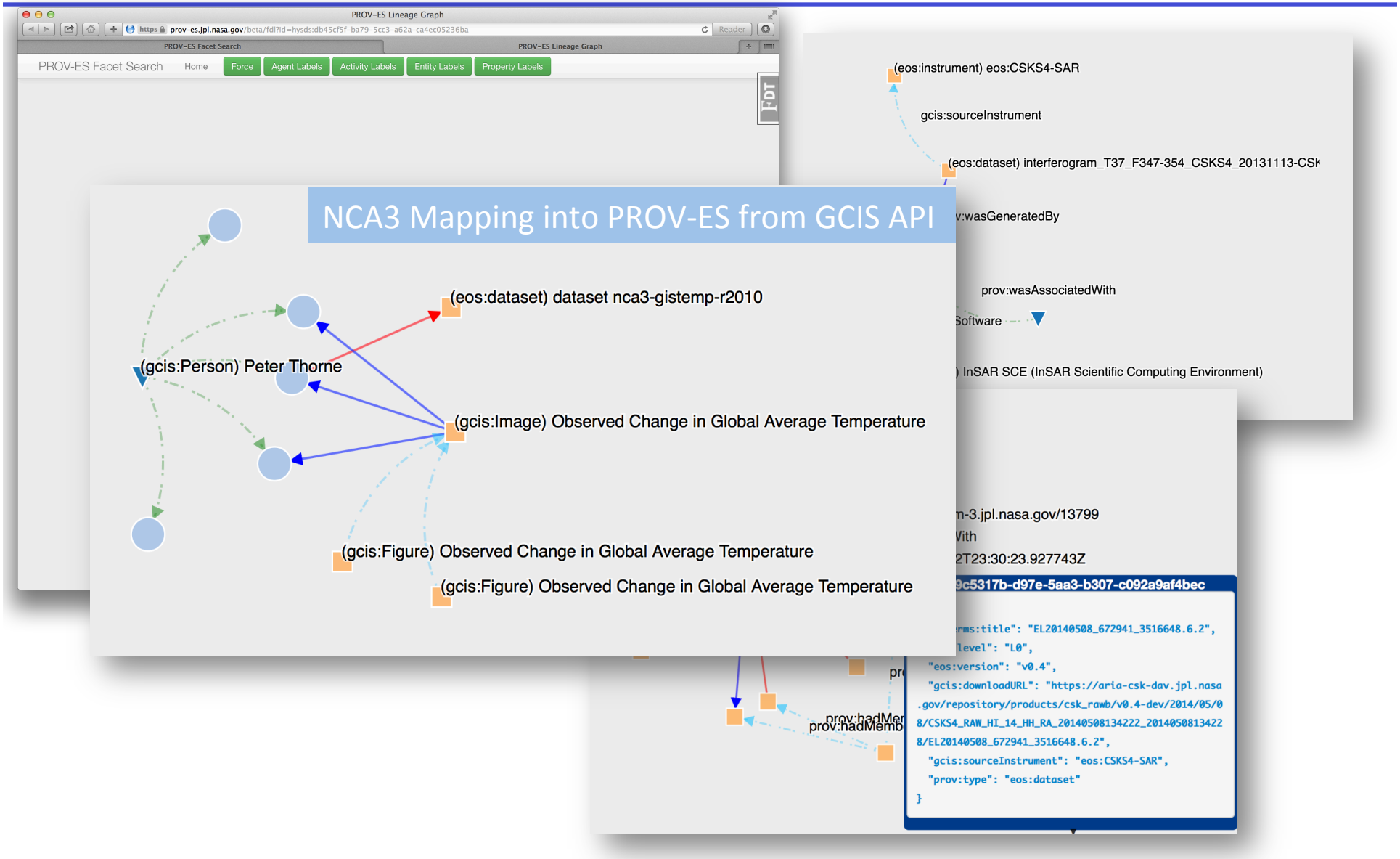
# PROV-ES Faceted Search



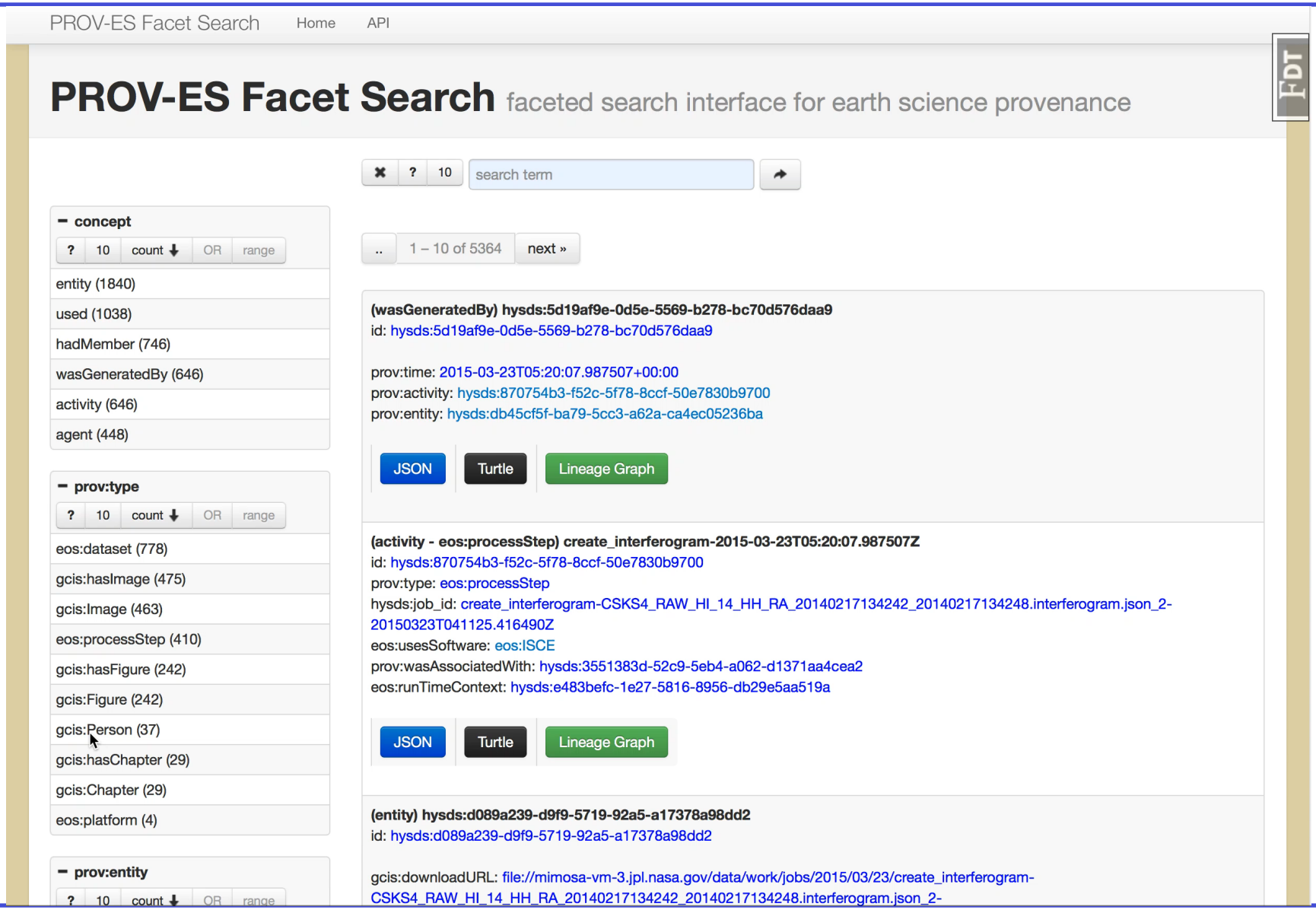
# Linking Provenance Concepts


- Support real-world scenarios of **independent** provenance records
  - Status quo of limited use of coordinated end-to-end **identifiers and provenance bundles**
- Support loosely-coupled and **linked data** provenance resources
- **Ingest** distributed provenance traces
- Dynamically assemble provenance bundles
  - Generate a **reachability graph** from any given provenance resource
  - Via identifiers, with fallback to search terms

# Interactive Visualization for Linked Relationships



# Interactive Visualization for Linked Relationships



# Recommendations


1. Coherently integrate results from other Data Stewardship Interest Area (DSIA) working groups into PROV-ES
   - Preservation Information Architecture WG (2013)
   - Data Preservation Practices (DPP) WG (2014)
   - Digital Object Identifier (DOI) WG (2013 & 2014)
   - Data Quality (DQ) WG (2014)
2. Initial PROV-ES ready for further infusion, feedback, and improvements
   - E.g. additional projects, DAACs, ESDIS, AIST, ACCESS, MEaSUREs, etc.
   - Setup community-wide instance(s) of PROV-ES Faceted Search for more usage
     - Benefits from linked data across seemingly incoherent provenance records
   - Can package up PROV-ES Faceted Search into (re)deployable virtual machines
3. Open up approach to broader communities. For example:
   - ESIP Data Stewardship Committee
   - ESIP Semantic Web Committee
   - ISO 19115 Lineage ? ISO 19115:2003, ISO 19115-2:2009, ISO 19115-1:2014
   - BEDI
   - CDI
   - GCIS

# Resources

- PROV-ES Faceted Search and API
  - https://prov-es.jpl.nasa.gov/beta/
- ESDSWG PROV-ES Working Group
  - Wiki
    - https://wiki.earthdata.nasa.gov/display/ESDSWG/Provenance+for+Earth+Science+(PROV-ES)+Working+Group (login required)
  - Mailing list
    - esdswg-prov-es@lists.nasa.gov
    - Subscribe at https://lists.nasa.gov/mailman/listinfo/esdswg-prov-es
  - Telecon notes
    - https://docs.google.com/document/d/1mw0ROVcW4-7W3IfZk2pizM1k6sHM2hgmRTvChrYtpO8/edit?pli=1
  - Use Case Mapping
    - https://docs.google.com/spreadsheet/ccc?key=0AlQ95ca89UmYdEJXX05NVDRhRGRWV09hcmNENnFROHc#gid=0
  - PROV-ES Mapping to W3C PROV
    - https://docs.google.com/spreadsheets/d/1gsyuKY-xD1yQaHD6QVB6vJ3YmDWL9WKPiSopW8BtuB8/edit#gid=0
- NASA Earth Science Data Preservation Content Specification (NASA ESD PCS)
  - http://earthdata.nasa.gov/sites/default/files/field/document/423-SPEC-001_NASA%20ESD_Preservation_Spec_OriginalCh01_0.pdf
- Provenance and Context Content Standard
  - http://wiki.esipfed.org/index.php/Provenance_and_Context_Content_Standard

# Thank You!


- Contributing members 2013-2014
  - Hampapuram Ramapriyan
  - Curt Tilmes
  - Brian Wilson
  - Gerald Manipin
  - Mary Jo Brodzik
  - Stace Beaulieu
  - Brian Duggan
  - Bob Downs
  - Helen Conover
  - Anne Wilson
  - Stephen Olding